국어의 어휘연구를 위한 통계 언어학의 원리와 방법

신 익 성
(서 울 大 學 校)

차 례

F. de Saussure 는 집단속에 존재하는 언어의 통계학적인 근원을 지적했다. 그는 어휘에 관해서 다음과 같이 말했다: ≪모든 어휘는 그 근원에 있어서 개인적인 창조이라는 것은 두말할 여지가 없으나 특히 집단적인 창조이다. 개인이 창조한 낱말은 개인이 받아들이고서 반복하는 정도안에서만이 그 가치를 가진다. 또한 낱말은 결국 그 사용의 총화에 의하여 한정된다. 사용은 사용의 전체에 있어서 언어의 상태를 반영하는 사용이다.≫

위에서 인용한 것은 특히 어휘가 통계학적인 근원을 더욱 더 많이 가지고 있다는 뜻으로 해석된다.[1]

언어현상의 모든 과학적인 기술, 언어현상에 관한 모든 결론은, 얻어진 자료의 통계학적인 취급을 정도의 차이는 있지마는 고려하지 않을 수 없다. 사전에서 주어진 단어의 뜻 문법에서 발견된 규칙은 재료의 체계적인 집대성을 모으는 동안에 혹은 언어 사용을 매일 경험하는 동안에 얻어진 많은 관찰의 평균가치이다. 전통 음성학에서 발견된 생리학적인 기술이 평균가치이고 청각음성학의 결과도 평균가치이다. 사적언어학도 음성법칙 혹은 언어사이의 친족관계를 세우는 데 있어서 많은 자료의 수집을 전제로 삼았고 발견된 수가 결론을 결정했다. 저작자를 결정하고 택스트(text)의 방언을 결정하는 데 있어서도 문헌학자들은 상위

[1] P. Guiraud, Problémes et méthodes de la statistique linguistique, P.U.F., Paris, 1960, p. 19

한 문헌이 가지고 있는 여러가지의 언어특징의 빈도수비교에 많이 의지했다.[2]

언어의 양적인 연구는 본질적으로 빈도의 관념에 의지하고 있다. 모든 언어기호, 모든 낱말은 집단안에서의 사용의 빈도수가 나타내는 윤곽의 총화와 함께 화자의 머리안에 있다. 그리고 개인은 그가 이 기호에게 주는 빈도에 의하여 집단속에서 언어상태에 이바지한다. 화자의 말에서 나타나는 낱말의 빈도는 결코 개인의 말의 우연한 사실뿐만이 아니라 그 낱말의 의미론적인 가치만큼 그 낱말의 내재적인 성질이다. 텍스트의 수적인 구조의 총화 혹은 같은 시대에 살고 있는 여러 개인들의 담화의 수적인 구조의 총화는 그 시대의 언어상태의 표현이라고 생각될 수 있다.

위에서 언급한 언어학의 많은 분야에서 자료의 수적인 취급은 더욱 더 중요한 역할을 하고 있다. 예를 들면 실험음성학에서 치조음의 지속 강도 주파수 억양곡선과 같은 것들에 있어서 평균치와 변동범위를 결정하는 문제가 더욱 더 중요시되고 있다. 오늘날의 음성학자들은 하나 혹은 두개의 개체에서 나온 각개 요인의 약간의 측정으로서는 만족하지 않고 있다. 한저작자의 언어를 문체론적으로 기술하기 위해서도 학자들은 그의 언어습관의 개산적인 평가로써는 만족하지 않고 확고한 근거를 빈도수 편차 평균치에 두려고 한다.

정보이론 분야에서의 업적중에서 통계학적인 계산과 수학적인 방법에 의지하지 않은 것은 없다. 음성학자들은 언어의 기술이 소리의 형의 목록 (inventory)과 한 형을 다른 형으로 부터 구별하는 특징의 기술 및 그들의 결합을 지배하는 특징의 기술에서 정지해서는 안된다는 의견으로 기울어지고 있다 언어의 기술은 우리가 가지각색의 형의 상대적인 빈도의 표적과 이 형들이 어휘안에서 뿐만 아니라 기록된 말 (speech) 혹은 인쇄된 텍스트안에서 일어날 때 이 형들이 어떻게 결합하느냐에 대한 표적을 가질 때까지는 완전하지 못하다. 질적인 기술은 양적인 기술에 의하여 보충되어야 된다. 지속되는 말안에서 개개의 음소가 어떻게 일어나느냐에 대한 지식이 중요한 만큼 더욱 더 큰 요소의 구성을 결정하는 법칙에 대한 지식이 중요하다. 우리는 낱말구조분석의 일부로서 낱말안의 음절의 수를 조사할 수 있다. 근년에 낱말 빈도에 대한 연구가 많은 언어를 대상으로 하여 활발하게 진행되고 있다. 이 연구는 이론적 실용적 중요성을 가지고 있다. 필자는 이 논문에서 어휘연구와 관련된 언어통계학만을 취급하고자 한다. 언어통계학의 원리 및 방법을 빈도와 분배 (repartition)과 관련시켜서 논하고 빈도 분배와 함께 기본어휘를 정하기 위한 세개의 주요한 기준중의 하나인 자유사용 (degré de disponibilité)에 관해서 진술하고자 한다.

1. 빈 도

만약 국어국정교과서 6권안에 있는 모든 낱말의 수가 16424이고 이중에서 3409가 동사

[2] Bertil Malmberg, New trends in linguistics, Stockholm, 1964 p. 187

이고 이 동사중에서 45가 〈가다〉라는 동사라면 동사의 절대빈도수는 3409이고 〈가다〉의 절대빈도 수는 45이다. 동사의 상대적빈도수는 3409/16424=0.2075이고 〈가다〉의 상대적빈도수는 45/16424=0.00274 이다. 동사중에서의〈가다〉의 상대적빈도수는 45/3409=0.0132 이다.

2. 어휘의 범위

텍스트가 주어지면 먼저 이 텍스트를 구성하고 있는 단어를 반복과 관계없이 센다. 이렇게 해서 우리는 이 텍스트의 크기의 측정인 수적인 가치 N를 얻을 것이다. 다음에 우리는 N안에 있는 상이한 낱말(Vocables)의 수 V를 얻을 것이다.

만약 텍스트안에서 동일한 단어가 두번 나타난다면 N는 둘이지마는 V는 하나이다. V는 N의 함수이다. 정해진 텍스트에 있어서 V는 N와 함께 늘어난다. 그러나 V는 N보다는 적게 늘어난다. 텍스트의 시작에서는 V는 어떤 낱말이 두번 반복하기 전에는 N와 같다. 반복이 증가함에 따라서 앞에서 한번도 나타나지 않은 낱말을 만나는 것은 더욱 더 드물게 될 것이다. 그러나, N는 같은 비율로써 늘어나는 것을 중지하지 않는다.

어휘의 범위(수V)는 텍스트의 범위(수N)의 함수이다. 만약, 우리가 충분히 동질이라고 생각된 텍스트에서 길이가 다른 두개의 발췌(fragment)를 얻는다면 긴 것이 짧은 것보다 큰 범위의 어휘를 가지게 된다. 그러나, 우리가 대단히 동질이라고 생각된 같은 길이의 여러개의 발췌를 얻었더라도 각각의 발췌의 어휘의 범위 안에는 편차가 있을 것이다. 여기서 문제가 되는 것들은 통계학적인 법칙을 따르는 우연적인 편차와 평균치 및 평균치에 대한 평균편차이다. 평균편차는 편차의 자승을 평균함으로써 얻어진다. 이 편차가 통계학적인 법칙이 예측하는 편차보다 크다면 이 편차는 문체론적 사실에 의한 것이라고 판단을 내릴 수 있다.

[N에 의한 V의 질적인 평가]

텍스트 A: $N=14217$

$V=1653$

우리는 실제로 N와 함수관계에 있는 V의 수학적인 기대를 계산할 수 없다. 숫자로 표시된 어휘의 범위가 우리의 기대 즉 평균치를 능가하면 그 어휘는 풍부하다고 평균치보다 적으면 빈약하다고 판단을 내릴 수 있을까? 그러나 이러한 판단은 다만 상대적인 것에 지나지 않는다. 이 판단은 다른 텍트스와 관련시키지 않고서는 성립될 수 없다.

텍스트 B: $N=16424$

$V=1536$

우리는 텍스트 A가 B보다 풍부한 어휘를 가지고 있다는 결론을 내릴 수 있다. 텍스트

A의 크기를 N_a, B의 크기를 N_b로 표시하고 각기어휘의 크기를 V_a, V_b로 표시하여

$$N_a \leqslant N_b \quad V_a > V_b$$
$$N_a < N_b \quad V_a = V_b$$

라는 공식을 얻는다면 A의 어휘가 B의 어휘보다 풍부하다고 말할 수 있다.

	A	B	A—B
N	14256	16690	−2434
V	1715	1518	+197

N와 V에서의 편차가 모두 +이고 N에서의 차이가 V에서의 차이보다 적다면 A의 어휘가 B의 어휘보다 풍부하다고 말할 수 있다.

$$N_a > N_b \quad V_a > V_b \quad (N_a - N_b) < (V_a - V_b)$$

A	B	차이
14256	14217	+39
1715	1633	+62

만약 이 조건이 충족되지 않는다면 어휘의 평균빈도 $\bar{f} = N/V$를 개입시킬 수 있다. 어떤 텍스트가 더욱 더 길고 평균빈도가 더욱 더 낮으면 그 텍스트의 어휘가 더욱 더 풍부하다고 말할 수 있다.

$$N_a > N_b \; V_a > V_b \; \bar{f}_a < \bar{f}_b$$

	A	B	편차
N	17173	16669	+504
V	1498	1355	+143
$\bar{f}$	11.464	12.302	−0.838

그러나 모든 차이가 방향이 같다면 즉 모든 차이가 모두 + 혹은 −이면 이 방법으로써 결론을 내리는 것은 불가능하다.

	A	B	편차
N	16677	17611	−934
V	1667	1687	−20
$\bar{f}$	10.00	10.44	−0.44

또 한편으로는 이 비교방법은 두 텍스트의 길이가 대단히 불균등하다면 이용가치가 없다.

	A	B	차이
N	2438	1447	+991
V	690	494	+196
$\bar{f}$	3.53	2.93	+0.60

3. 빈도의 분포(distribution)

빈도분류의 등급

우리는 동일한 빈도를 가지고 있는 V를 모아서 빈도의 등급을 얻을 수 있다. 각기 등급은 하나의 지수(effectif)이다. V_i는 빈도 f_i를 가진 V의 수라면 다음과 같은 공식이 성립될 수 있다.

$$V=V_1+V_2\cdots+V_k \quad 혹은 \quad V=\sum_1^k V_i$$

$$N=1V_1+2V_2\cdots+kV_k \quad 혹은 \quad N=\sum_1^k iV_i$$

어휘분포의 불변수(constants)

빈도가 늘면 지수(effectifs)는 준다. 한 텍스트안에서 빈도 f의 V가 100개 있다면 빈도 $f-1$의 V는 100개 이상이고 빈도 $f+1$의 V는 100개 이하이다. 또 한편으로는 가장 높은 지수를 가지는 것은 빈도 1이다. V_1는 언제나 V_2를 능가한다. 빈도가 증가하면 지수는 대단히 낮아져(1 혹은 대단히 적은 수의 단위) 0가 되기도 한다. 빈도사이의 간격이 수와 넓이에 있어서 증가함에 따라서 1이상인 지수는 더욱 더 드물게 된다. 우리는 N/V로써 $\bar{f}$를 V_1/V로써 p_1를 얻을 수 있다. p_1는 빈도 1의 V의 텍스트가 가지고 있는 V의 총화에 대한 비율이다. 어느 텍스트의 N가 16690이고 최대빈도가 885이라면 f_{max}/N는 0.053이다. $\bar{f}$ p_1 f_{max}/N 등을 어휘연구에 이용할 수 있다.

변 수(variables)

크기가 같지 않은 텍스트의 분포를 비교하면 우리는 다음과 같은 사실을 확인한다.

N가 늘면
- —빈도의 수가 는다.
- —최대빈도가 는다.
- —어떤 빈도의 지수가 는다.

어떤 텍스트가 문체론적으로 동질이다라는 가정을 세운다면 가장 높은 빈도를 가진 V는 텍스트의 크기에 현저하게 비례된 절대적인 빈도를 가진다. 우리말에서는 <이다>, 토씨등이 가장 높은 빈도를 가지고 있다. 적은 수의 어휘단위 혹은 문법단위가 모든 말(discours)의 대부분을 차지하고 있다. 관계를 표시하는 낱말이 불가결하고 빈번한 불어와 같은 언어에서는 빈도가 가장 높은 50개의 단위가 텍스트의 50% 이상을 차지한다. 가장 빈도가 높은 1,000개의 단위가 텍스트의 약 85%를 차지한다는 사실은 우리들로 하여금 일종의 기본어휘를 작성할 수 있는 가능성을 제공한다.[3] 텍스트 T가 가지고 있는 N와 V(V의 지수

[3] G. Gougenheim, P. Rivenc, R. Michéa, A. Sauvageot, L'élaboration du français fondamental, Paris, 1967. 참조.

$V_1, V_2, \cdots, V_k$와 함께)를 안다면 N'를 가진 텍스트 T' 안에 있는 최대빈도 K'에 대한 이론적가치를 계산할 수 있다. 그러나 우리가 지금까지 관찰한 것으로는 V', V_1', V_2'에 대한 계산은 불가능하다.

우리는 길이는 같으나 어휘의 수가 다른 텍스트에서 얻어진 분포를 비교할 수 있다. 더욱 풍부한 어휘를 가진 텍스트에서 우리는 다음과 같은 사실을 확인한다.

—표시된 빈도의 수가 일반적으로 더욱 낮다.

—최대빈도가 일반적으로 더욱 낮다.

—V_1가 더욱 많다. 만약 텍스트가 상당히 길다면 이 차이는 V_2에 대해서도 역시 나타난다. 전형적인 가치사이에서 성립되는 관계를 생각한다면 텍스트가 길어지는 경우에는 우리는 다음과 같은 사실을 확인한다.

—평균빈도 즉 $\bar{f}=N/V$는 는다. V는 N보다 적게 증가한다.

—텍스트의 어휘의 총화에 대한 빈도 1을 가진 어휘의 비례 즉 $p_1=V_1/V$는 감소한다.

—두개의 연속된 빈도의 지수 f_i와 f_{i+1} 사이의 비례 혹은 V_i/V_{i+1} 사이의 비례는 준다. V_i는 V_{i+1} 보다 적게 는다.

크기는 같으나 어휘가 더욱 풍부한 텍스트에서는 다음과 같은 사실이 확인된다.

—평균빈도 $\bar{f}$가 더욱 낮다. ($N_a=N_b$이지마는 $V_a>V_b$이라는 사실에서부터)

—비례 p_1가 더욱 높다. 그러나 이것은 확실한 것은 아니다. $V_a>V_b$와 $V_{1a}>V_{1b}$가 있기 때문이다.

적어도 낮은 빈도에 대해서는 비례 V_i/V_{i+1}가 더욱 크다.

분포의 운동

텍스트의 시작에서는 어떤 V가 반복되지 않는 동안은 $V_1=V$이다. 따라서 $\bar{f}=1$이고 $p_1=1$이다. 그러나 반복이 일어나자 마자 V_2가 나타남으로써 V의 수는 N보다 적어지고 V_1의 수는 V보다 적어진다. 이 순간부터 $\bar{f}>1$ $p_1<1$이 된다.

두번 나타난 V의 수는 늘어나고 p_1는 준다. $V_3, V_4\cdots$이 나타남에 따라서 $\bar{f}$는 한없이 늘어나는 반면에 p_1는 0에 가까워진다. 어떤 V들이 빈도 1을 버리고 빈도 2로 빈도 2를 버리고 빈도 3으로 감에 따라서 V_1와 같은 새로운 V들이 텍스트안에 나타난다. 이 새로운 V들은 잃은 것보다는 얻는 것이 많다.

빈도 1에서 빈도 2로 가는 V의 수는 빈도 2에서 빈도 3으로 가는 V의 수보다 많다.

주의해야 될 점이 셋이다.

1) 언어의 정상적인 행사만이 통계연구의 대상이 되어야 한다. 우리는 동일한 낱말을 열번 반복함으로써 말(discours)을 시작할 수 있다. 이야기의 어떤 길이 동안은 어떤 V도 반복하지 않고 말을 할 수 있다. 우리는 이렇게 해서 괴상스러운 분포를 만들 수 있을 것이다. 하나는 $V=1$ $N=10$ $V_1=0$이고 다른것은 $V=N$ $\bar{f}=1$이다. 이것은 고려할 만한 대상이 못

된다.

2) 너무 짧은 표본은 피해야 된다. $N=10$에 대한 V의 평균치를 찾는다는 것은 문제거리가 되지 않는다. N가 적어도 100을 넘어설 때 조사연구가 의미를 가지게 된다. N가 약1000이 될 때 비로소 문체론적인 불변수(constantes)가 나타난다.

3) 이법칙들은 상당히 많은 수에 적용될 때 비로소 가치를 가진다. 어떤 제한 이하에서는 우연한 것이 일어날 수 있다. 어떤 표본안에서 체계적인 반복이 일어날 수 있다. 이렇게 되면, V_1는 붕괴되고 V_2, V_3…은 팽창한다. 따라서 P_1는 약해진다. 이와 반대로 새로운 V가 지나치게 늘어나면 f가 내려가고 P_1이 상승한다. 그러나, 이러한 현상은 일시적인 것이다.

4. Zipf의 법칙

Zipf는 빈도×순위(rang)는 불변수이다라는 법칙을 발견했다. 빈도가 가장 높은 낱말은 순위 1이고, 그 다음 빈도가 높은 낱말은 순위 2이다. James Joyce의 소설 "Ulysses"에서 빈도순위가 10인 단어가 2,653번 사용되어 있고, 순위100인 단어가 265번 나타난다. 순위 1,000인 단어는 26번 사용되어 있다.

$10\times 2,653=26,530$

$100\times 265=26,500$

$1,000\times 26=26,000$

이 계산법을 Corneille의 희극 rôle d'Alcandre에 적용하니 다음과 같은 결과가 나타났다. 가장 빈도수가 높은 1부터 5까지와 빈도수 1(가장 낮은 것)을 제외한 나머지 모든 것의 f(빈도)$\times r$(순위)은 240과 320사이에 있다.[4] Zipf법칙의 장점은 텍스트의 크기와 거의 관계가 없고 또한 사용된 언어(idiome)와도 관계가 없다는 데 있다. 그러나 이 법칙이 적용되는 경우에 예외없이 꼭 같은 차이가 나타났다. 이 법칙은 가장 높은 빈도에는 효력이 없고 가장 낮은 빈도에 적용될 때는 거의 언제나 불규칙성이 나타난다. 이 법칙은 이상과 같은 단점을 가지고 있지마는 현대의 정보이론에 의하야 설명되는 그리고 기억의 구조와도 관계가 있는 어휘의 오묘한 경향을 표현하고 있다. Zipf는 $f\times r=$constante에 의하여 재미나는 가설을 세웠다. 그는 이 균형을 두개의 대립된 힘사이에서 발견된 조화라고 본 것이다. 말은 두개의 경쟁에 의하여 지배된다. 말하는 사람은 내용에 의하여 요구되는 정확한 낱말 대신에 가능한 한 동일한 낱말을 반복하는 경향을 가지고 있다. 그 반면에 듣는 사람은 사용된 낱말로써 상세하게 기술하고 최대한의 다양성을 표현하는 것과 함께 최대한의 투명을 요구한다.

[4] Charles Müller, Initiation ä la statiatique linguistique, Paris, 1968, p. 166

"표현될 모든 개념을 위한 동일한 낱말"과 "개념의 각기를 위한 특수한 낱말의 두극단 사이에는 위에서 언급된 방정식에 의하여 표현된 균형이있다. 이 균형은 실제에 있어서 최소한의 노력의 원리이다. Guiraud는 또한 낱말의 길이와 빈도사이에 존재하고 Zipf의 법칙과 동일한 경향으로 돌려야 될 또 하나의 재미나는 법칙을 세웠다. 가장 많이 사용되는 낱말이 가장 짧고 빈도가 낮은 낱말이 가장 많은 수의 음운을 가지고 있다. ($f \times r$=constant, 이방정식에서 f는 빈도이고 r은 기호가 요구하는 에네르기의 정도 즉 언어에 있어서는 음운의 수) 우리말에 있어서는 토씨는 내용의 영향을 거의 안받기 때문에 거의 모든 텍스트에서 상대적으로 변치 않는 빈도를 가지고 있다.

불란서 어휘연구소는 Zipf의 법칙에 의하여 다음과 같은 어휘일람표를 전자계산기를 사용하여 작성했다.

Rang r	Fréquence	Nombre	$f \cdot r\alpha$ pour $\alpha=1,305$	Total tranche	Fréquences cumulées	% de-compréhens
1	14,083	1	14,083	14,083	14,083	4.51
2	11,552	1	28,543	11,552	25,635	8.21
3	10,503	1	44,051	10,503	36,138	11.58
4	7,905	1	48,260	7,905	44,043	14.11
5	7,515	1	61,388	7,515	51,558	16.52
6	6,846	1	70,945	6,846	58,404	18.71
7	5,308	1	68,100	5,374	63,778	20.43
⋮	⋮	⋮	⋮	⋮	⋮	⋮
28	2,391	2	184,978	4,782	141,526	45.34
⋮	⋮	⋮	⋮	⋮	⋮	⋮
5,296	1	2,700	72,402	2,700	312,135	100.00

조사된 낱말의 총수 : 312,135. 상위한 낱말의 수 : 7,995

이 일람표에서 각 칸은

첫째 낱말의 순위

둘째 이 순위와 일치되는 빈도

셋째 동일한 빈도를 가진 낱말의 수

넷째 $\alpha=1,305$ $f \times r\alpha$=constante

다섯째 $f \times$(동일한 빈도를 가진 낱말의 수)

여섯째 각기 빈도 계단에 이것보다 높은 빈도 계단 총수를 합친 것

일곱째 이해의 퍼센티지 즉 조사된 낱말의 총수를 fréquence cumuleé로 나눈 것.[5]

한 낱말의 빈도P_1 P_2 P_m를 발견하는 것이 언제나 가능하다면 m개의 가장 빈도가 많은

[5] 참조 Bulletin d'information du laboratoire d'Analyse Lexicologique, no. 6, Publication du Centre d'Etude du Vocabulaire Français de la Faculté des Lettres et Sciences Humaines de Besançon.

단어의 총빈도를 찾아내는 것 즉 어떤 수의 가장 빈도가 많은 단어가 텍스트의 몇%를 차지하느냐를 찾아내는 것은 언제나 가능하다.

$$\sum_{r=1}^{m} Pr = \sum_{r=1}^{m} kr - \gamma$$

$k=0.1$ $\gamma=1.01$ $m=1,100$ 이라고 가정한다면 우리는 $\sum_{r=1}^{m} Pr=0.8$를 얻을 수 있다. 즉, 1,100개의 가장 많이 사용되는 낱말들이 텍스트의 80%를 차지한다.[6]

5. Waring-Herdan의 공식

$$\underset{(1)}{\frac{x-a}{x}} + \underset{(2)}{\frac{(x-a)a}{x(x+1)}} + \underset{(3)}{\frac{(x-a)a(a+1)}{x(x+1)(x+2)}} + \cdots + \underset{(n)}{\frac{(x-a)a(a+1)\cdots(a+n-1)}{x(x+1)(x+2)\cdots(x+n)}} = 1$$

(2)는 (1)보다 적고 (2)는 (3)보다 적다. 이 분포는 왼쪽에서 바른쪽으로 각기 항(terme)이 감소되는 형식으로 되어 있다. 그리고 이 항들의 총화는 1이다. 각기 항을 결정하는 변수는 x와 a이다.

$$a=\frac{1}{\frac{1}{q^1}-\frac{1}{\bar{f}}-1}, \qquad x=\frac{a}{q^1}$$

$\bar{f}=N/V$ $P_1=V_1/V$ (이미 앞에서 정의된 것들이다). $q_1=(1-P_1)$ q_1는 V개의 낱말중의 하나가 1보다 많은 빈도를 가지는 확률이다.[7]

만약 이 공식이 잘 적용된다면 n번째의 항은 텍스트중의 한 단어가 빈도 fn를 가지는 확률을 제공해야 된다.

$$P_n=\frac{(x-a)a(a+1)\cdots(a+n-2)}{x(x+1)(x+2)\cdots(x+n-1)}$$

$N=1629$, $V=534$ $V_1=366$ (Corneille, Rôle d'Alcandre의 어휘)

0.685	0.136	0.057	0.030	0.019	0.013	0.009	0.007	0.0055	0.0044…	(+0.0341)
1	2	3	4	5	6	7	8	9	10	>10

$=1$

$V=534$ $534\times0.685=365.790$ 366

$534\times0.0136=72.624$ 72.6

[6] O.S. Akhmavova, I.A. Mel'chuk, R.M. Frukina, E.V. Paducheva, Exact methods in linguistic research, University of California Press, 1963 p. 106

[7] Gustav Herdan, Quantitative Linguistics, Londres, 1964 p. 85

366 는 빈도가 1 인 단어의 이론적 수(effectif theorique)이고 72.6 는 빈도가 2 인 단어의 이론적 수이다.

f_i	V_i(이론적)	V_i(실재의)	편차	x^2
1	(366)	366		
2	72.6	71	−1.6	2.56/72.6=0.04
3	30.4	28	−2.4	5.76/30.4=0.19
4	16.1	15	−1.0	1.00/16.0=0.06
5	10.1	12	+1.9	3.61/10.1=0.36
6	6.9	5	−1.9	3.61/ 6.9=0.52
7	4.8	5	+0.2	0.04/ 4.8=0.01
8−15	16.6	12	−4.6	21.16/16.6=1.27
>15	10.6	20	+9.4	88.36/10.6=8.34
	534.0	534	0.0	x^2=10.79

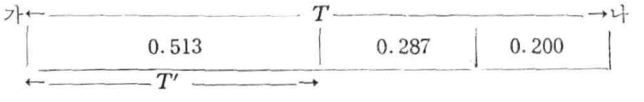
빈도가 낮은 것은 편차가 작지마는 빈도가 높은 것은 편차가 크기 때문에 이 공식에 의하여 가장 좋은 결과를 얻는다는 것은 어려운 듯이 보인다.

N가 100,000를 넘어서면 이 공식을 적용해서는 안된다. 우리말의 기본어휘일람표를 만드는 연구에는 이공식은 적용될 수 없다. 그러나 우리나라의 작가의 작품안에 나타나는 어휘연구에는 적용될 수 있다. 이것도 대단히 낮은 빈도를 가진 낱말에 한한다.

6. 어휘의 이론적범위

A 라는 작품이 3 부 (a, b, c) 로 되어 있다고 가정하자. 만약, 우리가 A 의 어휘에 관한 모든 기지수(旣知數)를 가지고 있다면, a, b, c 중의 어떤 하나의 이론적 어휘를 계산할 수 있다.

	N		V	f	N'/N
A	3177		808	3.93	
a	1629		534	3.05	0.513
b	913	합계 3177	328	2.78	0.287
c	635		272	2.33	0.200



우리가 534, 328, 272 라는 실제의 수와 가까운 수를 계산해 낼 수 있다. 추첨을 하는 경우에 어떤 것이 T' 안에 있는 확률은 p=0.513 이고 T' 안에 있지 않는 보조확률은 q=0.487

이다.

T 안에서 한번만 나타나는 단어의 수가 488이라고 가정하고 488이 a, b, c 사이에 아무렇게나 분배되어 있다면 T' 안에 나타나는 이론적인 수는 $0.513 \times 488 = 250.3$이고 T' 안에 있지 않은 수는 $0.487 \times 488 = 237.6$이다.

T 안에 두번 나타나는 단어의 수는 132이다. 132가 T' 안에 나타나는 확률은 p^2이다. T' 밖에 나타나는 확률은 q^2이다. 이 확률을 계산하는데 쓰이는 공식은 $(x+y)^2$이다.

f_2 T' 안에서	$0.513 \times 0.513 = 0.2632$	34.7
f_1	$2 \times 0.513 \times 0.487 = 0.4996$	66.0
O	$0.487 \times 0.487 = 0.2372$	31.3
	1.0000	132.0

T' 안에 한번도 나타나지 않은 단어의 수는 다음 공식에 의하여 계산될 수 있다.

f_i	V_i	q^i	$q^i V^i$
1	488	$0.487q^1$	$237.6\ q^1V_1$
2	132	$0.237q^2$	$31.3\ q^2V_2$
3	56	$0.1155q^3$	$6.5\ q^3V_3$
4	33	$0.0562q^4$	1.9
5	24	$0.0274q^5$	0.7
6	6	$0.0133q^6$	0.1
			278.1

즉 $E(V_0') = q^1V_1 + q^2V_2 \cdots + q^kV_k$

$E(V_0') = \Sigma q^i V_i$

$E(V_0') = 278.1$

T가 가지고 있는 단어의 합계는 808이므로 T' 안에 있는 단어의 이론적인 수는 $808 - 278.1 = 529.9$이다. 실제의 단어수는 534이다. 따라서 편차는 $+4.1$이다. 이 편차는 큰 것이 못된다.

7. 어휘의 이론적 구조

(1) 이론적 effectifs의 계산(낮은 빈도) T' 안에 없는 단어의 수 V_0'를 계산하는 데 사용된 이항식(二項式)이 0이외의 낮은 빈도를 계산하는 데도 사용된다.

f	$f'=0$	1	2	3	4	5	6
1	$1q$	$1p$					
2	$1q^2$	$2qp$	$1p^2$				
3	$1q^3$	$3q^2p$	$3qp^2$	$1p^3$			
4	$1q^4$	$4q^3p$	$6q^2p^2$	$4qp^3$	$1p^4$		

5	$1q^5$	$5q^4p$	$10q^3p^2$	$10q^2p^3$	$5qp^4$	$1p^5$	
6	$1q^6$	$6q^5p$	$15q^4p^2$	$20q^3p^3$	$15q^2p^4$	$6qp^5$	$1p^6$
⋯	q^i	$C_i^1q^{i-1}p$	$C_i^2q^{i-2}p^2$	$C_i^3q^{i-3}p^3$	$C_i^4q^{i-4}p^4$	$C_i^5q^{i-5}p^5$	$C_i^6q^{i-6}p^6\cdots C_i^{i-1}qp^{i-1}$

$f'=0 \quad E(V_0')=\Sigma q^i V_i$

f 는 빈도이고 f' 는 준빈도(sous-fréquence)이다.

준빈도 1 $\quad E(V_1')=\Sigma_i pq^{i-1}V_i$

준빈도 $f_j' \quad E(V_j')=\Sigma C_i^j p^j q^{i-j} V_i$

a 와 b 안에 있는 V_1' 를 계산하는데 2 항식을 적용하면 다음 표와 같은 결과가 나온다.

V_i	A	$a(p=0.513)$	적(積)	$C(p=0.200)$	적(積)
1	488	1×0.513	250.3	1×0.2	97.6
2	132	$2\times0.513\times0.487$	65.9	$2\times0.2\times0.8$	42.2
3	56	$3\times0.513\times0.487^2$	20.4	$3\times0.2\times0.8^2$	21.5
4	33	$4\times0.513\times0.487^3$	7.8	$4\times0.2\times0.8^3$	13.5
5	24	$5\times0.513\times0.487^4$	3.4	$5\times0.2\times0.8^4$	9.8
6	6	$6\times0.513\times0.487^5$	0.5	$6\times0.2\times0.8^5$	2.4
7	8	$7\times0.513\times0.487^6$	0.4	$7\times0.2\times0.8^6$	2.9
8	8	$8\times0.513\times0.487^7$	0.2	$8\times0.2\times0.8^7$	2.1
9	5	$9\times0.513\times0.487^8$	0.2	$9\times0.2\times0.8^8$	1.2
10	5			$10\times0.2\times0.8^9$	1.1
11	2			$11\times0.2\times0.8^{10}$	0.4
12	4			$12\times0.2\times0.8^{11}$	0.7
13	4			$13\times0.2\times0.8^{12}$	0.6
14	1			$14\times0.2\times0.8^{13}$	0.1
		V_1'(실제의 편차)	349.1		196.1
			366		194
			+16.9		−2.1

a 와 c 를 비교하면 c 가 편차가 작다.

a 는 편차가 크므로 어휘가 c 에 비해서 풍부하다.

(2) 이론적 빈도의 계산(평균빈도와 높은 빈도)

우리가 평균빈도 혹은 높은 빈도에 다다르면 확률이 감소되기 때문에 a, b, c 안에 있는 이론적인 effectif 가 적은 분수(分數)가 될 정도로 effectifs 이 감소된다. 이렇게 되면 2 항식을 effectifs 에 적용하지 않고 빈도에 적용하는 것이 좋은 방법이다.[8]

f_i	a 이론적	실제	c 이론적	실제
218	111.8	117	53.6	50
202	103.6	107	40.4	44

[8] 참조 Charles Müller, Initiation à la statistique linguistique, Paris, 1968, p. 180-181

94	48.2	52	18.8	19
65	33.3	17	13.0	9
65	333.3	40	13.0	11
etc.				

8. 빈 도 〔1〕

우리가 텍스트의 문체론적 특징을 알기 위해서 어휘를 양적으로 분석할 때 혹은 텍스트안에 있는 어휘의 구조를 통하여 각자의 어휘를 평가하려고 할 때 무엇보다도 중요한 것은 빈도 1의 낱말이다.

텍스트의 범위에 비해서 어휘의 범위가 크다면 이 어휘가 빈도 1의 낱말을 많이 가지고 있다면 그 이유가 어디 있는지를 살펴 보아야 된다. 텍스트 A를 같은 크기의 a, b로 나누었을 때 b보다 a안에 빈도 1의 단어가 많다면 이 텍스트를 쓴 사람이 a에서 상황낱말(lexique de situation)을 많이 썼다는 것으로 생각된다.

일반적으로 텍스트의 시작에서는 거의 모든 단어의 빈도는 1이다. 그러나 텍스트의 시작에서는 사용되지 않았던 단어들이 텍스트안에 나타나면 이것은 우리에게 의의있는 사실을 제공할 가능성을 가진다. 빈도 1의 단어는 문체론적 요소로써 취급되어야 되고 상황단어가 풍부하다는 것과 관계가 있다고 생각되어야 된다.

만약 우리가 텍스트를 다섯개의 부분으로 나누었다고 가정한다면 텍스트 전체안에 빈도 1이 몇개 있는지와 빈도 1이 다섯개의 부분사이에서 어떻게 분배되어 있는지를 조사해야 된다. 분배(repartion)에 관한 것은 다음 장에서 논하겠다. 만약 텍스트가 부분으로 나누어졌을 때 부분의 크기가 같다면 비교는 대단히 쉬운 일이다. 그러나 부분의 크기가 같지 않으면 빈도 1의 수의 차이가 문체론적 차이에 의한 것인지 혹은 부분의 같지 않은 크기에 의한 것인지를 판단하기 어렵다.

비교에 의한 방법

$$N_a > N_b$$

$$V_a > V_b$$

$$V_{1a} < V_{1b}$$

위의 공식에 의하면 텍스트 B가 텍스트 A보다 어휘가 풍부하다.

	C	D	
N	16677	17611	+934
V	1667	1687	+20
V_1	718	704	−14

C가 D보다 어휘가 풍부하다.

다음과 같은 예에서는 어느 쪽이 풍부하다는 것을 알기가 조금 까다롭다.

	E	F	
N	16690	16501	−189
V	1518	1493	−25
f	10.995	11.052	+0.057
V_1	548	550	+2

189/16690=0.011

548×0.011=6.028

E와 F의 크기를 동등하게 한다면 E가 잃는 것은 6.028이다.

1518−6=1512는 1493보다 많음으로 V_1에 있어서 편차가 조금 있음에도 불구하고 E의 어휘가 F보다 풍부하다는 것을 인정할 수 있다.

9. 낱말의 분배(repartition)

앞장에서 설명된 계산방법은 텍스트 안에 있는 낱말의 완전한 검사를 가정했다. 빈도분포 일람표를 가정했다.

한 낱말 혹은 낱말의 무리가 텍스트의 부분들 사이에서 어떻게 분배되어 있는지를 조사할 필요가 있다. 이 분배를 이론적인 분배와 비교하는 것 또는 약간의 낱말들을 규칙성에 의하여 분류하는 작업 또는 규범부터 많이 벗어난 것들을 따로 떼어놓는 것은 대단히 중요한 조사이다. 규범부터 많이 벗어 난 것은 문체론적인 문제와 관계가 깊다.

x^2 혹은 Peason의 검사

이론적가치를 C로써 실제의 가치를 O로써 표시한다면 검사의 일반적 공식은 다음과 같이 된다.

$$x^2=\Sigma\frac{(O-C)^2}{C}$$

어떤 텍스트안에서 어떤 문법적요소가 과잉이냐 아니냐를 판단하는데 이 공식이 사용된다. 이것을 판단하는데 있어서 기준이 되는 것은 이론적 가치이다. Corneille의 희극의 배역중의 하나인 Matamore 안에 있는 낱말의 총수는 2438이고 명사의 수는 484이다. 그러나 희극전체 안에는 명사가 18%이다. 따라서 이론적인 명사의 수는 $2438\times\frac{18}{100}=438$이다.

	이론적가치 C	실제의 가치 O	편차 $(O-C)$	x^2 $(O-C)^2/C$
명사……	438	484	+46	2116/438=4.831
다른낱말……	2000	1954	−46	2116/2000=1.058
	2438	2438	0	x^2=5.889

우리는 X^2의 일람표에 의하여 5.889를 해석한다. 언어통계학에 관한 책에는 X^2의 일람표가 있다. 438+2,000=2,438에서 2,438은 이미 알려진 수임으로 2,000과 438중에서 하나만 알면 다른 것을 알 수 있다. 이것을 자유도(degré de liberté)이라고 부른다. X^2의 일람표에 의하면 X^2=5,889는 0.01과 0.02사이이다. 무가설(hypothèse nulle)이 사실이라면 우리가 X^2=5,889를 적중시킬 수 있는 기회는 $\frac{1}{100}$과 $\frac{2}{100}$사이이다. 따라서 우리가 무가설을 거부함으로써 속는 기회는 $\frac{1}{100}$과 $\frac{2}{100}$사이이다. 무가설과 반대되는 가설은 문체론적인 의미를 가지고 있는 변이(variation)의 가설이다. 문체론적인 현상이 실재로 존재한다고 결론을 내릴 수 있다.

우리는 어떤 텍스트를 크기가 같은 부분으로 나누고서 이부분들안에 나타나는 두개의 다른 낱말을 x에 의하여 비교할 수 있다. 우리말의 어떤 텍스트안에 있는 (은,는)과 (이,가)를 x에 의하여 비교하여 이 둘중에서 모두 다 정상적인지 혹은 과잉인지 둘중의 하나가 과잉인지에 대한 판단을 내릴 수 있다.

어떤 텍스트를 세개의 부분으로 나누고 이 부분들의 크기가 같지 않을 때는 각각의 이론인 수를 세어야 된다.

		이론적	실재의	편차	x^2
(1) 텍스트	a 0.603	570.5	555	−15.5	240.25/570.5=0.42
	b 0.205	193.9	173	−20.9	436.81/193.9=2.25
	c 0.192	181.6	218	+36.4	696.96/181.6=3.84
	1.000	946.0	946	+0.0	6.51
(2) 동일한 텍스트	a	205.5	198	−7.0	49.00/205.0=0.24
	b	69.7	77	+7.3	53.29/69.7=0.76
	c	65.3	65	−0.3	0.09/65.3=0.00
					x^2=1.00

(1)은 불어의 정관사를 (2)는 부정관사를 x에 의하여 조사한 것이라고 가정한다면 자유도는 2이므로 가치 1은 무가설로써 적중되는 확률이 50% 이상이고 가치 6.51은 적중되는 확률이 0.05이하이다. 이 두 낱말은 분배가 정규적이냐 아니냐가 문제가 될 정도로 다르다.

산포지수(L'indice de dispersion)

N가 500,000인 텍스트를 크기가 같은 5개의 부분으로 나눈다. 어떤 한 단어의 빈도가 f이라면 5개의 하위빈도(sous fréquences)는 f'_1, f'_2, …, f'_5이다. 하위빈도의

평균치 $\bar{f}'=\frac{f}{n}$

평균편차 $\sigma=\sqrt{\sum\frac{(f'_i-\bar{f}')^2}{n}}$

여기서는 n=5이다.

변이계수(coefficients de variation)

$v=\frac{\sigma}{\bar{f}'}$ 이다.

만약 그 낱말이 완전하게 규칙적으로 분배되어 있다면 편차, 평균편차, v는 모두 다 0이다. 그 낱말이 5개의 부분중의 한부분안에서만 나타나면 이 분배는 가장 불규칙한 것이 된다. 그렇게 되면 하위빈도는 $f, 0, 0, 0, 0$이다. v는 극한 $\sqrt{n-1}$에 다다른다. 여기서는 $\sqrt{5-1}=2$이다. 만약 부분이 10개라면 극한은 $\sqrt{10-1}=3$이다. 지수(indice)를 0과 1 사이에 두기 위해서는 v를 $\sqrt{n-1}$로 나누면 된다.

$$0\leqslant\frac{v}{\sqrt{n-1}}\leqslant 1$$

이 공식과는 반대로 분배가 가장 불규칙이면 0이 되고 가장 규칙적이면 1이 되는 D로서 표시되는 산포지수는

$$D=1-\frac{v}{\sqrt{n-1}} \qquad 0\leqslant D\leqslant 1$$

다음은 불어의 부정사 ne의 산포지수이다.

f'_i	$f_i'-\bar{f}$	$(f_i-\bar{f})^2$
2439	+972	944784
1439	−28	784
1323	−144	20736
1226	−241	58081
906	−260	313600
7334	−1	1337985

평균치($\bar{f}$) : 1466.8(수를 늘이면 1467($5\times 0.2^2=0.2$의 정정)

$(\bar{f}'_i-\bar{f})^2=133984.8$

$\sigma=\sqrt{\frac{1337984.8}{5}}=517.3$

$v=\frac{517.3}{1466.8}=0.3527$

$D=1-\frac{0.3527}{2}=0.8237$

이 D의 값은 분배가 대단히 규칙적이라는 것을 가리킨다.

$x^2=\frac{1337984.8}{1466.8}=912.2$

x^2의 값은 이 편차가 우연이 아니라는 것을 가리키고 있다.[9]

[9] 참조 M.A. Juilland, Dictionaire de fréquence du français, Mouton.

10. 어휘의 관계

발간된 날짜 문체 주제까지도 거의 비슷한 동일한 저자에 의하여 쓰여진 두개의 텍스트를 비교할 수 있다. 이와 반대로 저자 주제 문체가 다른 두개의 텍스트의 비교도 가능하다. 동일한 작품일지라도 이것을 두개의 텍스트로 나누어서 텍스트의 어휘를 서로 비교하면 대단히 다른 결과를 얻을 수 있을 것이다.

구조와 어휘의 내용

두개의 텍스트가 구조는 같으나 어휘의 내용은 현저하게 달라질 수 있다. 반대도 가능하다 어휘의 관계라는 말을 내용에만 적용하는것이 적당하다고 생각된다. 구조의 비교는 이미 진술한 방법에 의하여 조사될 수 있다. 두개의 텍스트의 크기가 같다면 두개의 빈도분배일람표를 비교하는 것은 직접적이다. 두개의 텍스트의 크기가 같지 않으면 우리는 더욱 큰 자료에 의하여 작은 텍스트를 위해서 실재의 분배와 대조될 이론적인 분배를 계산할 것이다. 비교는 문법범주의 분배 혹은 어휘의 다른 범주로 확장될 수 있다. 그러나 낱말의 동일은 고려하지 않는다.

어휘의 관계를 위한 무가설

어휘내용에 관해서는 먼저 무가설을 명확하게 해 두어야 된다. 두개의 텍스트의 어휘가 같다는 것이 무가설이다. 전체 A+B를 만들기 위해서 텍스트A와 B를 결합하면 이 전체의 어휘는 두개의 텍스트중의 한 텍스트 안에서만 나타나는 빈도1의 정해진 수의 낱말을 포함하게 된다. 빈도2 빈도3도 마찬가지로 된다. 빈도2는 이중의 $\frac{1}{4}$은 A 안에서만 나타나고 $\frac{1}{4}$은 B 안에서만 나타난다. 빈도3은 $\frac{1}{9}$은 A 안에서만 나타나고 $\frac{1}{9}$은 B 안에서만 나타난다. 낮은 빈도에 있어서는 각기 텍스트는 다른 텍스트안에 없는 일정한 수의 낱말을 가질 것이다. 이 자체도 두개의 텍스트가 최대한의 어휘적관계를 가지고 있을 때만이 가능하다. 두개의 텍스트로써 만들어진 전체의 어휘적 구조가 정규적이고 부자연스럽지 않다는 조건밑에서만이 가능하다. 전체의 어휘가 두개의 하위전체사이에서 우연적으로 분배되어있는 것이 무가설이다. 모형 (modéle)은 두개의 각기 텍스트의 크기와 그들의 어휘의 구조에 의존한다. 실제로 관찰된 것과 무가설사이의 편차를 일으키게 하고 어휘적관계를 감소시키는 조건은 생략하겠다.

어휘관계의 측정

두개의 텍스트의 어휘가 공통된 것을 가지고 있지 않을 때 0이고 그들의 관계가 최대한인 경우에 0가 되는 어휘관계를 측정할 수 있는 관계지수(indice de connection)가 있다.

텍스트A의 N_a=1447 텍스트B의 N_b=1458 텍스트의 크기는 비슷하다. 그러나 A의 V는 494, B의 V는 451이라면 어휘의 크기에 있어서는 상당히 큰 차이가 있다. 두개의 텍스트

를 결합하면 V는 729이고 729중의 216은 두 텍스트에 공통된 것이고 278은 A만이 가지고 있는 V이고 235는 B만이 가지고 있는 V이다.

	A	$\bar{A}$	총 계
B	216	235	451
$\bar{B}$	278	—	278
총 계	494	235	729

216과 729과의 관계를 $CV_{(ab)}$로써 표시함으로써 다음과 같은 공식을 얻는다.

$$CV_{(ab)}=\frac{216}{729}=0.296$$

각기 텍스트의 독립을 다른 것에 비교하여 측정할 수 있다.

$$\bar{C}V_{(a\to b)}=\frac{278}{494}=0.563 \qquad \bar{C}V_{(b\to a)}=\frac{235}{451}=0.521$$

다음은 두 텍스트에 공통된 낱말의 수의 관계를 표시하는 지수이다.

$$r=\frac{216}{\sqrt{494\times 451}}=\frac{216}{472}=0.458$$

텍스트의 크기의 영향을 받는 것이 이 지수의 약점이다. 대단히 작은 두개의 텍스트(단지 약간의 낱말)를 비교하면 관계지수는 0이 되고 독립지수는 1(하나)가 된다. 텍스트를 늘임에 따라서 관계지수는 증가하고 독립지수는 감소된다. 이 지수들이 어떤 최대한과 어떤 최소한으로 향하느냐를 말하기는 어렵다. 또 하나의 불편한 것은 이 측정이 두개의 텍스트 사이의 크기의 관계에 의하여 영향을 많이 받는다는 것이다. 하나가 대단히 크고 또 하나가 대단히 작다면 두 텍스트에 공통된 낱말의 수는 더욱 작은 텍스트의 어휘의 범위와 거의 같게 될 것이다. 독립지수는 하나는 0이 될 것이고 다른 것은 1에 대단히 가까울 것이다. 같은 크기의 텍스트에 대해서만 이 지수를 참작하는 것이 신중한 태도이다.[10]

위에서 진술한 것은 어휘연구에 필요한 언어통계학의 원리와 방법이었다. 빈도와 분배가 중심문제가 되었던 것이다.

빈도와 분배만으로는 우리말의 기본어휘연구가 불가능하므로 이 연구에 반드시 필요한 문제를 다음에 논하고자 한다.

11. 구체적일 낱말과 빈도

일반적으로 빈도표에는 구체적인 낱말이 거의 없다. 이 구체적인 낱말들은 대체로 빈도가 낮다. 예를 들면 구체적인 낱말중의 하나인 이〔齒〕라는 낱말은 이가 아프다라는 경우들을

[10] 참조 Charles Müller, Etüde de statistique lexicale, Paris, 1967, p. 169.

제외하고는 거의 사용되지 않고 있다. 우리가 이를 사용하지 않은 날은 하루도 없지마는 이 낱말의 빈도는 대단히 낮다. 빈도는 낮지마는 중요한 낱말이라는 것은 부인할 수 없다. 게다가 구체적인 낱말의 빈도는 안정성이 적다. 외국의 빈도와 분배가 중심이 된 사전을 보면 우리는 분배지수에 의하여 이 구체적인 낱말들이 안정성이 없다는 것을 알 수 있다. 이것은 모든 언어에 공통된 경향이다. 왜냐하면 구체적인 낱말은 어떤 상황 대화의 어떤 제목에 묶여 있기 때문이다. 어느 사람이 치과병원을 나오는 친구를 만날 때 그 친구는 이라는 말을 사용할 것이다. 일반적으로 문법적인 관계를 나타내는 낱말, 동사등이 빈도가 높고 안정성도 많다. 이러한 낱말들은 언어의 형식 혹은 틀이 되는 것들이고 구체적인 낱말은 언어의 내용이 되는 것이다. 이라는 낱말은 중요한 낱말이므로 빈도가 낮고 안정성이 적다는 것은 다만 소극적인 성질이다. 이러한 낱말들의 적극적인 성질을 세워놓는 문제가 남아 있다. Renè Michéa 씨는 Langues modernes 이라는 논문에서 mots ≪thématiques≫와 mots ≪athématiques≫를 구별했다. Mickéa 씨의 정의에 의하면 mots athématiques 는 텍스트의 내용과는 관계없이 충분한 크기의 어느 텍스트에서나 거의 규칙적으로 나타나는 낱말이다. 이것들은 사물 자체를 표현하는 데보다도 차라리 사물의 주제에 대하여 표현하는 데 이바지하는 낱말들이다. 따라서 모든 주제 모든 상황에 다소 공통된 낱말들이 mots ≪athématiques≫이다. 이와 반대로 mots ≪thématiques≫는 주어진 어떤 종류의 제목에 묶여 있다. 이 낱말들은 대상 자체를 표시한다. 이 낱말들은 대체로 구체적인 낱말들이다. 이 낱말들의 빈도는 조사된 텍스트의 선택에 의존한다.

mots ≪athématiques≫와 mots ≪thématiques≫와의 구별은 교육적인 견지에서는 옳고 중요할지라도 이 구별은 문제를 전적으로 해결하지 못한다. 특히 이 구별은 mots thématiques 를 빈도가 높은 낱말과 대비하여 특징지우고 있지 않다. 자주 쓰이는 낱말(mots fréquents)과 자유롭게 쓸 수 있는 낱말(mots disponibles)을 대립시키는 것이 더욱 잘된 구별이다.

mots disponibles 은 대화가 경해진 제목을 취급하는 경우를 제외하고는 자주 사용되지 않지마는 필요하면 자유롭게 사용될 수 있기 때문에 우리는 그와 같이 부르는 것이다. 이 낱말들은 순식간에 잊어지지도 않고 실어증(aphasie)에도 완강하게 버티는 낱말들이다.[12]

12. 자유로 사용할 수 있는 낱말과 그 낱말의 자유 사용 가능도 (degré de disponibilité)

빈도의 통계외에서도 이 두 어휘(자주 쓰이는 낱말과 자유로 사용할 수 있는 낱말)의 구별은 대단히 명확하다. 자유로 사용할 수 있는 낱말만이 우리들의 관심을 야기시킨다. 빈도가

[11] R. Michéa, Vocabulaire et Culture, Les Langues Modernes, 1950, pp. 188-189
[12] G. Gougenheim, L'Elaboration du Français Fondamental, Paris, 1964, p. 145

높은 낱말은 비록 유용할지라도 그렇지 못하다.

Michéa 씨는 프랑스 Périgueux 고등학교서 Gougenheim 교수는 프랑스의 Héléne-Boucher 고등학교에서 학생들에게 다음과 같은 질문에 대답하라고 요구했다.

≪여러분들이 지금부터 기차여행을 한다고 가정하십시오. 지금 여러분들은 역에 있읍니다. 먼저 머리에 떠오르는 20개의 낱말을 쓰십시오.≫ 이 질문에 대한 회답이 의미심장했다. monter 를 제외하고는 학생들이 구어에서 빈도가 높은 대단히 일반적인 동사를 인용하는 것을 소홀히 했다. 동사가 소홀히 된 반면에 명사가 많이 인용되었다. 우리는 이와 같은 종류의 조사에서 주요한 구체적인 명사가 현저한 안정성을 가지고 나타났다는 것을 인정할 수 있다. 어떤 상황이 주었을 때는 동사가 명사보다 불안정하다는 것은 일반적인 현상이다. 이 조사의 재미나는 면은 한편으로는 조사된 각기 어휘의 분산의 비교이고 또 한편으로는 각기 일람표의 내부에서 문법적인 성질에 따라서 분배를 연구하는 것이다. 조사에서 인용된 낱말의 총수에 대한 각기 일람표의 다른 낱말의 수의 비례가 분산(dispersion)이다.

분산의 비교

Périgueux 고등학교의 일람표 : 인용된 총수 400 에 대하여 149 개의 다른 낱말 37.86%

Hélène-Boucher 고등학교의 일람표 : 700 에 대한 279 개의 다른 낱말 39.86%

Périgueux 의 표와 Héléne-Boucher 의 표에서 명사가 다른 품사에 비해서 압도적으로 많이 나타났다는 결과가 나왔다. 어떤 상황이 주어졌을 때 가장 먼저 마음에 떠오르는 낱말은 특별히 이 상황과 관계가 있고 이 상황을 특징지우는 낱말들인데 이 낱말들은 명사이다. 빈도표에서는 동사가 가장 안정성이 있고 명사가 안정성이 없는데 관심의 중심주위에 있는 관념의 연합에 대한 조사에서는 구체적인 명사가 안정성이 있고 동사가 안정성이 없는 것으로 나타났다. 우리에게 가장 필요한 어휘는 빈도가 높은 어휘와 자유로 쓸 수 있는 어휘를 합친 것이다. 자유처분의 계단이라는 관념이 이 낱말들이 우리 기억안에서 다소 직접적으로 존재한다는 것과 일치한다. 자유처분계단이라는 관념은 기본어휘의 방법론적인 확립을 위해서 대단히 중요하다. 자유처분계단을 결정하기 위해서는 관심중심의 방법에 의지해야 된다. Michéa 는 Langues Modernes 이라는 논문에서 16 개의 관심중심표를 작성했다.

1) 신체의 부분

2) 의복

3) 가옥

4) 가구

5) 음식과 식사의 음료

16) ——

Michéa 가 정한 이 16 개의 관심의 중심이 인간의 모든 관심을 포함하지는 못한다. 관심중심은 년령, 성별, 직업, 생활양식에 따라서 달라지지마는 우리는 가능한 모든 인간에 공

통된 관심중심 일람표를 작성하도록 노력해야 된다. 국민학교의 상급학년생과 중고등학교 학생들이 조사의 좋은 증인이 될 수 있다.

예를 들면 우리는 어떤 학교의 50명(이수는 일정한 것이 아니다)의 학생들에게 다음과 같은 카드를 배부하여 학생들이 기입한 단어를 조사할 수 있다.

관심중심 의복

1) 성명
2) 학교이름
4) 소년 혹은 소녀
5) 학년
7) 날짜
8) 부모의 직업

9) 20개단어표
1. ……
2. ……
3. ……
4. ……
5. ……
⋮
20. ……

우리나라의 도(道)마다 한 학교씩 선택하여 의복에 관한 조사에서 얻어진 20개의 단어를 비교하면 우리는 지리적인 원인에 의한 차이가 있는지 어떤지를 알 수 있을 것이다. 프랑스에서 조사한 결과에 의하면 그차이는 대수롭지 않았다. 이것은 모든 언어에 공통된 현상이기 때문에 우리말도 예외가 아니라도 생각된다. 관심중심의 구체적인 낱말이 지리적인 원인과는 관계없이 그들의 자유처분 계단에 따라서 거의 동일한 순서로써 마음에 떠오른다면 이것은 심리적인 현상이 현저하다는 것을 의미한다. 이와같은 안정성은 빈도의 일반적인 일람표에는 없다. 지리적인 원인에 의한 차이가 발견되면 우리는 그 차이를 설명해야 된다. 그 차이는 대체로 설명하기 쉽다. 일반적으로 이 지방적차이는 근소하며 전체의 결과의 가치를 위태롭게 하지는 않는다.

관심의 중심에 관해서 농촌과 도시를 비교하면 대단히 유익한 결과가 나올 것이다. 우리는 소년과 소녀와의 비교에서는 도시와 농촌의 비교에서 만큼 흥미있는 자료를 얻지못할 것이다. 프랑스에서 조사한 표에 의하면 시골의 소년 소녀들은 주요한 낱말에 그들의 투표를 집중시키는 반면에 도시의 소년 소녀들은 세목에 그들의 주위를 분산한다. 따라서 자유처분계단은 도시에서보다는 시골에서 훨씬 더 높다. 이것도 불어에만 국한된 현상이 아니고 모든 언어에 공통된 것이라고 생각된다. 소년과 소녀와의 사이의 자유처분의 비교는 의의 깊은 차이를 보여 주지 않는다.

13. 자유사용에 관한 고찰

우리는 도(道)별로 조사된 일람표를 비교할 수 있다.

A 도

관심의 중심(신체의 부분)			
1	86	1	86
2	83	1	83
3-4	81	2	162
5	73	1	73
⋮	⋮	⋮	⋮
		총계 5	

B 도

관심의 중심(신체의 부분)			
1	91	1	91
2-3	87	2	174
4	86	1	86
5	79	1	79
⋮	⋮	⋮	⋮
		총계 5	

위의 표에서 첫째란은 낱말의 선순을 둘째란은 얻어진 투표의 수(자유처분계단)를 셋째란은 이 자유처분계단을 가진 낱말의 수를 넷째란은 둘째와 셋째의 적(積)을 표시한 것이다. 셋째란 아래쪽에 쓰여진 숫자는 관심의 중심을 위하여 인용된 낱말의 수(dispersion)이다. 이렇게 비교해 보면 동일한 관심중심을 위한 각기도의 곡선은 현저하게 평행한 선으로 나타난다. 이 평행선은 동일한 관심 중심에 대한 반응이 유사하다는 것을 의미한다. 위의 표에서 최대자유처분계단은 100이다.[13] 위에서 정의한 바와 같은 자유처분의 원칙은 빈도가 뚜렷이 들어나게 할 수 없는 구체적인 낱말을 제공한다. 또 한편으로는 관심의 중심은 어휘의 전체를 포함하지 못한다. 관심중심의 방법은 비록 유효하지마는 우리가 대단히 중요하다고 생각하는 모든 교육적인 어휘를 제공할 수 없다. 우리가 기본어휘를 정하기 위해서는 빈도분배, 자유사용계단이라는 세개의 기준을 고려해야 된다. 어떤 낱말이 빈도 자유사용계단에 있어서는 대단히 중요한 낱말이라고 판단되지마는 분배가 5 이하라면 기본어휘에서 제외되어야 되느냐 혹은 제외되어서는 안되느냐라는 문제가 생긴다. 신중히 다루어져야 될 문제이다.

빈도가 몇 이하는 제외되어야 되느냐 자유사용계단 몇 이상은 고려되어야 되느냐를 결정하는데는 원칙이 있어야 되고 충분한 이유가 있어야 된다. 한가지 부언하고 싶은 것은 조사를 위한 증인은 직업, 연령, 지방, 성별에 따라서 광범위하게 망라되어야 되고 직업이 없는 부녀도 좋은 증인이 될 수 있다. 조사하는데 있어서의 섬세하고 미묘한 기술문제는 언어지리학에 있어서의 기술문제와 비슷하다.

---

[13] G. Gougenheim, L'Elaboration du Français Fondamental, Paris, 1964, pp. 189-194

맺 는 말

어휘연구와 관련된 통계의 원리와 방법 그리고 기본어휘조사를 위한 세개의 기준 즉 빈도분배, 자유사용계단에 관한 것을 논했다. 특히 방법과 원리를 논하는데 있어서 중심이 된 것은 정확하게 정의된 무가설에 일치하는 표본의 구조와 이 표본과 실제로 관찰된 것사이의 편차의 평가였다. 우리가 어떤 영역안에서 양적인 자료를 수집했을 때 특히 이 자료가 큰 덩어리가 되었을 때 일어나는 문제를 해결하기 위해서는 통계적인 고찰이 절실히 요구됨에도 불구하고 우리나라 학자들은 이 고찰을 너무도 소홀히 하고 있는 듯하다. 언어통계학에 대한 지나친 무관심 때문에 우리나라에는 구미의 거의 모든 나라가 가지고 있는 기본어휘 일람표마저 없다. 통계언어의 원리와 방법은 비교적 간단하다. 그러나 실제와 가까운 확률을 얻기 위해서는 많은 표본이 필요하게 된다. 조사되어야 될 단위의 양이 방대하기 때문에 연구기구(예를 들면 전자계산기) 없이는 조사연구가 거의 불가능하다. 언어의 통계학적연구를 위해서는 통계수학자와 언어학자가 서로 협조해야 된다. 통계수학자는 이론적 표본을 만들기 위한 추리와 계산의 방법과정을 제공하고 관찰된 편차의 확률을 측정하기 위한 검증을 담당한다. 그밖의 모든 것은 언어학자의 소관이다. 관찰된 사실을 결정하고 이 사실에 어떤 의의를 부여한다는 등 결과를 해석하는 것은 언어학자의 일이다.

우리말의 기본어휘 조사연구는 우리말의 연구와 교육을 위해서 시급한 문제이다. 여기서 논한 원리와 방법은 모든 언어에 적용될 수 있는 성질의 것이지마는 이 논문의 목적은 추상적이나마 우리말의 기본 어휘의 조사연구를 위한 원리와 방법을 제시하는 데 있다. 실제의 조사를 위해서는 언어학자, 통계학자, 국어학자들이 협동연구조사를 해야 된다.

<Synopsis>

The principles and methods of the statistic linguistics for the lexical study of Korean

Ik Sung Shin

F. de Saussure indicated the statistical origin of language which is in the linguistic collectivity. The general meaning of the indication is as follows: all vocabulary in its origin is of an individual creation, especially collective creation; the words an individual creats have

its values only in the extent that an individual accepts and repeats; furthermore a word is defined by the total sum of its usage; its usage is the usage which reflects the situation of language in its entire usage.

Above mentioned Saussure's indication can be interpreted to the effect that vocabulary especially has statistical origin more than the other linguistic units.

All the scientific description of linguistic facts and all the conclusion about them presuppose the statistical treatment of the given materials. The meanings of words given in dictionaries and the rules found in grammars are in actual fact the mean values of a number of observations which have been made in the course of our everyday experience of linguistic usage. The physiological descriptions found in traditional phonetics are mean values, so are the results of acoustic phonetics.

Even historical linguistics, in setting up "sound laws" or relationships between languages, presupposed the collection of a mass of data, and the number found decided the conclusions to be drawn. In deciding the authorship and dialect of text, philologists relied mostly on the comparative frequency of various linguistic features in the differents manuscripts: The quantitative research of language relies on the idea of frequency in its nature. All linguistic signs, all words are stored in the speaker's brain with the total sum of the features that the frequency of its usage in the linguistic collectivity represents; and an individual contributes to his linguistic situation in the collectivity by means of the frequency he gives to this signs.

The frequency of the words represented in the speaker's utterance is not only the incidental facts of the speech of an individual, but also so much the immanent nature of the words as the semantic values of the words. The total sum of the numerical structure of texts or of discources of individuals is the representation of the linguistic situation of the epoch.